Flexible & Executable Provenance in Data-Intensive Biomedical Research: A Flexible Research Data Service

DUKE UNIVERSITY

PI: HUANG, ERICH S.

Grant Number: 1 U01 EB020957-01

Tracing the lineage of scientific data and assertions is critical for the ""checks and balances that once ensured scientific fidelity"" (Collins & Tabak, 2014). As data become pervasively digitized, generating and following lineages automatically and at scale increases the usefulness and quality of conclusions. A significant challenge in data-intensive science is generating that lineage-the provenance-of scientific information, while facilitating retrieval and re-execution. We hypothesize such capabilities improve the reproducibility of assertions and make data more useful to society. Our objective is to build application programming interfaces for provenance, data-integrity, storage, and reproducible workflows that empower researchers to record, retrieve, and re-run scientific lineages. The rationale for the proposed research is that the value of the scientific data is enhanced by being able to retrospectively reproduce a result and by understanding its origins for future use. Provenance also facilitates measurement of data's importance-its impact. Guided by strong preliminary work, we will test our hypothesis by pursuing two specific aims: (1) Building APIs for provenance, data management, data integrity, and re-executable workflows, (2) Providing a platform for storing and deploying containerized compute environments that also serves as a learning laboratory for reproducible data science. This approach is innovative in focusing on flexibility and accommodating the myriad use cases across biomedical science, while pro- viding a hub for training investigators in reproducible data science. By creating an open source ""Flexible Research Data Service"", the proposed research will significantly impact our ability to make our investments in biomedical research more useful. PUBLIC HEALTH RELEVANCE PUBLIC HEALTH RELEVANCE: The proposed research is relevant to public health by developing an array of services that help researchers generate more robust and reproducible data-intensive science. It is relevant to the NIH's mission of supporting the application of knowledge to reduce the burdens of human disability.